AFRL-OSR-VA-TR-2014-0103

TOPOLOGICAL METHODS FOR DATA FUSION

Gunnar Carlsson
LELAND STANFORD JUNIOR UNIV CA

04/30/2014
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTC
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
31-03-2014	Final	07-01-2009 - 12-31-2013

4. TITLE AND SUBTITLE

Topological Methods for Data Fusion

5a. CONTRACT NUMBER

5b. GRANT NUMBER

FA9550-09-1-0531

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Professor Gunnar Carlsson

e-mail: gunnar@math.stanford.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Stanford University
3160 Porter Dr., Ste. 100
Stanford, CA. 94304-8445

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Office of Science and Research
875 Randolph Street
Suite 325 Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release: Distribution Unlimited

13. SUPPLEMENTARY NOTES

None.

14. ABSTRACT

The project has studied a number of questions related to data fusion. One is the investigation of zig-zag persistent homology, which is a tool which permits one to study the behavior of qualitative shape invariants (such as homology) across different samples from a larger data set, as well as understand time varying data sets. We have also demonstrated the utility of the Mapper methodology in the study of various different data types, by demonstrating that it provides insight into complex data sets that allow one to obtain useful knowledge concerning the data. In addition, we have constructed a coordinatization of the space of persistence barcodes, which will be very useful in allowing the application of traditional machine learning to data sets where the entries themselves carry geometric structure, such as chemical compounds. We have also applied earlier topological findings to construct compression schemes as well as algorithms for discriminating between types of texture patches.

15. SUBJECT TERMS

Topological data analysis, persistent homology, texture analysis, topological mapping, zigzag persistence

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Professor Gunnar Carlsson
U	U	U	UU	2	**19b. TELEPHONE NUMBER** *(include area code)* (650) 723-2224

Final Report on AFOSR Award FA9550-09-1-0531, *Topological Methods for Data Fusion*

We have pursued a number of projects in Topological Data Analysis. They concern two aspects of the subject, namely the "measurement of shape" via homological signatures and the compressed representation of the shape of a data set. Homological signatures from conventional topology have been extended to the notion of *persistent homology* [18], which is applicable to finite samples from spaces, perhaps with error, rather than only to the complete spaces themselves. The properties of persistent homology are under intense development, and part of this project constitutes work in that direction. We have also developed additional applications of persistent homology in the study of evolution of viruses. Compressed representations of point cloud data sets in the form of simplicial complexes are constructed using the *Mapper* methodology [4], [15], and in this project we have developed applications of the methodology as well as additional theory describing the stability of the construction.

1. **Zig-zag persistence:** Persistent homology is a methodology which has been developed to infer the topological properties of data sets, properly understood. The translation of point cloud data, which consists of discrete sets of points with an associated distance function, into geometric objects is done via a number of ways, including the *Vietoris-Rips* complex. An important feature is that persistent homology tracks the behavior of these complexes across a range of values of a threshhold parameter, and that the actual geometric features (as opposed to smaller features which represent noise) are those which persist (or are stable) across a large range of threshhold values, as that threshhold value increases. The complexes increase with the parameter, and therefore there are induced maps between them. There are, however, a number of situations where one would like to track values and assess consistency of a homological invariant over a set of values for which the complex is not increasing with a scale parameter. One example of this is where one selects a large number of samples from a very large data set. In this case, one wants to study persistence of features detected by homology across a range of samples rather than across the values of an increasing parameter. A second situation is where one is applying the so-called *witness complex* method [10]. Here one approximates the topology using a complex based on a set of *landmark points*, and to assess the fidelity of the construction it is useful to compare the topologies across different choices of landmarks. A third situation concerns time varying data, where one studies data sets for various time slices, which might overlap but in which neither is included in the other. A solution to these problems is provided by *zig-zag persistence* [5], [6]. This method has been studied in a number of situations during the course of the grant [17], with exploratory work confirming that it functions as an effective tool for assessing the consistency across samples, across landmarks, and across different choices of a variance parameter in kernel density estimators.

2. **Applications of Mapper:** In [15], a method was introduced for representing general data sets in compressed form as a simplicial complex. The idea is that a simplicial complex is an ideal representation for a data set, because by comparison with a scatterplot representation one obtains both compression of the structure as well as additional resolution, and by comparison with standard clustering methods one also obtains additional resolution. The method

had previously been used to study folding problems in molecular dynamics [2]. We have continued to demonstrate the value of the technique in several directions, including microarray studies of breast cancer, in the study of fragile-X (an autism related syndrome), and in politics and sports [12], [14], [11]. In the case of breast cancer, it permitted the identification of a genomic profile which characterizes a group of patients (roughly 8% of the patients) who all survived the length of the study. In the case of fragile-X, the finding was a decomposition of all the patients into two distinct groups, with distinct behaviors. The methodology makes it very simple to color the networks produced by variables of interest, making very clear the effects of various variables, which is very useful in understanding the relationships of the variables. The pertinence to data fusion comes from the fact that it is quite natural to apply the methodology to the set of "columns" attached to a data set, rather than the rows. This might mean the set of genes or gene sets used as coordinates in microarray studies, or it might mean the collection of sensors used in the study of geological data. The study of this kind of "sensor space" can detect many important pieces of information concerning the method of collecting data. For example, one might discover that a very large set of sensors is highly correlated. When this is the case, they will tend to dominate the geometry of the data set of rows, and one might want to compensate for this by dividing by the density in the column space to mitigate the problem.

3. **Texture analysis:** An earlier application of topological methods in data analysis was the application of persistent homology to identify the topological type of the space of high density high contrast 3×3 image patches in natural images [3]. Work on this award included an application of that finding to design a method for discriminating textures based on the topological structure of the space of patches. The methods relies on finding, for each high contrast patch in the texture patch, the closest point to the Klein bottle, and deriving from it a distribution of the Klein bottle. Fourier analysis is then carried out on the Klein bottle to obtain coordinates which discriminate well between patches from a benchmark data set of texture patches. One strong point of the method is the fact that the behavior of the coordinates under rotation of the patch is governed by a simple transformation law involving translation within the Klein bottle. It also suggest the possibility that understanding of the geometry and topology of frequently occurring patches in other imaging modalities should yield ways of understanding texture like situations in those modalities.

4. **Coordinates in bar code space:** Persistence bar codes have been shown to be useful for understanding the topology and geometry of individual data sets in [3], [16], and [8]. However, they can also be used in situations where the data points themselves are equipped with geometric structure. For example, databases of chemical compounds and of images have this property. In this situation, by assigning barcodes to the data points, we obtain a database of barcodes. Geometric structures on the collection of barcodes are in this case important for the purposes of analyzing the database, using, for example, methods from machine learning. One such structure is the *bottleneck distance* [9], which is a metric imposed on the set of all persistence barcodes. One might also attempt to find a coordinate system on the set of barcodes, for a more direct analysis. This idea has been explored and implemented during this project [1]. Specifically, we find that the set of all barcodes, with an equivalence relation which permits the deletion of any bars of length zero, can be described as a colimit of varieties, and has a ring of functions A of the form

$$A = \mathbb{R}[\tau_{ij}, 0 \leq i, 1 \leq j]$$

where τ_{ij} is the function which assigns to each barcode $\{[x_s, y_s]\}_s$ the sum

$$\tau_{ij} \sum_s (y_s - x_s)^i (y_s + x_s)^j$$

This idea can now be used to good effect to carry out the analysis of databases of chemical compounds, as has been demonstrated by other investigators. It also suggests a direction of attack on questions concerning multidimensional persistence. Multidimensional persistence is known not to possess a barcode description analogous to that for single variable persistence. However, it appears likely that this ring of functions can be extended to a ring of functions on sets of multidimensional persistence profiles. This would be extremely powerful, since it is by now clear to most investigators that multidimensional persistence profiles are of fundamental importance in the study of various kinds of data sets. A two dimensional profile based on both a scale or distance variable as well as a density variable would be extremely useful, and this work points in that direction.

References

[1] Adcock, A., Carlsson, E., and Carlsson, G., *The ring of algebraic functions on persistence bar codes*, arXiv 1304.0530, 2013.

[2] Bowman, Gregory R.,Huang, Xuhui, Yao, Yuan, Sun, Jian, Carlsson, Gunnar, Guibas, Leonidas J. and Pande, Vijay S., *Structural Insight into RNA Hairpin Folding Intermediates*, Journal of the American Chemical Society, vol. 130, 30, 9676–9678, 2008.

[3] Carlsson, Gunnar, Ishkhanov, Tigran, de Silva, Vin, and Zomorodian, Afra, *On the Local Behavior of Spaces of Natural Images*, International Journal of Computer Vision, vol. 76, 1, 1-12, 2008.

[4] Carlsson, Gunnar, *Topology and data*, Bull. Amer. Math. Soc. (N.S.) 46 (2009), no. 2, 255-308

[5] Carlsson, Gunnar and de Silva, Vin *Zigzag persistence*, Found. Comput. Math. 10 (2010), no. 4, 367-405.

[6] Carlsson, Gunnar, de Silva, Vin, and Morozov, Dmitriy. *Zigzag persistent homology and real-valued functions*, In Proceedings of the twenty-fifth annual symposium on Computational geometry (SCG '09). ACM, New York, NY, USA, 247-256, (2009).

[7] Carlsson, Gunnar and Zomorodian, Afra *The theory of multidimensional persistence*, Discrete Comput. Geom. 42 (2009), no. 1, 71-93.

[8] Chan, Joseph Minhow, Carlsson, Gunnar, and Rabadan, Raul *Topology of viral evolution*, Proc. Natl. Acad. Sci. USA 110 (2013), no. 46, 18566-18571.

[9] Chazal, Frederic, Cohen-Steiner, David, Guibas, Leonidas J, Mmoli, Facundo and Oudot, Steve Y., *Gromov-Hausdorff Stable Signatures for Shapes using Persistence*, Computer Graphics Forum, vol. 28, 5, 2009.

[10] De Silva,Vin and Carlsson, Gunnar, *Topological estimation using witness complexes*. In Proceedings of the First Eurographics conference on Point-Based Graphics (SPBG'04), Marc Alexa, Markus Gross, Hanspeter Pfister, and Szymon Rusinkiewicz (Eds.). Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 157-166, (2004).

[11] Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, Alagappan, M., Carlsson, J., Carlsson, G., *Extracting insights from the shape of complex data using topology*, Scientific Reports. 2013/02/07/online

[12] Nicolau, M., A. Levine, and G. Carlsson, *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival*, Proceedings of the National Academy of Sciences, (2011).

[13] Perea, Jose A., and Carlsson, Gunnar, *A Klein-Bottle-Based Dictionary for Texture Representation*, Int. J. Comput. Vis. 107 (2014), no. 1, 75-97.

[14] Romano, David, Nicolau, Monica, Quintin, Eve-Marie, Mazaika, Paul K., Lightbody, Amy A., Cody Hazlett, Heather, Piven, Joseph, Carlsson, Gunnar, Reiss, Allan L., *Topological methods reveal high and low functioning neuro-phenotypes within fragile X syndrome*, Human Brain Mapping, May, 2014.

[15] G. Singh, F. Memoli, and G. Carlsson, *Topological methods for the analysis of high dimensional data sets and 3D object recognition*, Point Based Graphics 2007, Prague, September 2007.

[16] G Singh, F Memoli, T Ishkhanov, G Sapiro, G Carlsson, DL Ringach, *Topological analysis of population activity in visual cortex*, Journal of vision 8 (8), 2008

[17] Tausz, A. and Carlsson, G., *Applications of zigzag persistence to topological data analysis*, arXiv 11.08.3545

[18] Zomorodian, Afra and Carlsson, Gunnar, *Computing persistent homology*, Discrete Comput. Geom. 33 (2005), no. 2, 249-274.